Spike to Spike Model and Applications: A biological plausible approach for the motion processing

Maria-Jose Escobar, Guillaume S. Masson, Thierry Viéville, Pierre Kornprobst

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Spike to Spike Model and Applications: A biological plausible approach for the motion processing

Maria-Jose Escobar — Guillaume Masson — Thierry Vieville — Pierre Kornprobst

N° 6280

September 2007

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--6280--FR+ENG

Spike to Spike Model and Applications: A biological plausible approach for the motion processing

Maria-Jose Escobar* , Guillaume Masson† , Thierry Vieville‡ , Pierre Kornprobst§

Thème BIO — Systèmes biologiques
Projet Odyssée

Rapport de recherche n° 6280 — September 2007 — 37 pages

Abstract: We propose V1 and MT functional models for biological motion recognition. Our V1 model transforms a video stream into spike trains through local motion detectors. The spike trains are the inputs of a spiking MT network. Each entity in the MT network corresponds to a simplified model of an MT cell. From the spike trains of MT cells a motion map of velocity distribution is built representing a sequence. Biological plausibility of both models is discused in detail in the paper. In order to show the efficiency of these models, the motion maps here obtained are used in the biological motion recognition task. We ran the experiments using two databases Giese and Weizmann, containing two (*march, walk*) and ten (*e.g. march, jump, run*) different classes, respectively. The results revealed that the motion map here proposed can be used as a reliable motion representation.

Key-words: spiking networks, motion analysis, V1, MT, biological motion recognition

* Maria-Jose.Escobar@sophia.inria.fr
† Guillaume.Masson@incm.cnrs-mrs.fr
‡ Thierry.Vieville@sophia.inria.fr
§ Pierre.Kornprobst@sophia.inria.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Spiking to Spike Model et Applications: Une modèle biologique pour le traitement du movement

Résumé : Pas de résumé

Mots-clés : réseau des neurones, traitement du movement, V1, MT, reconaissance du movement biologique

Contents

1 Introduction

Biological motion recognition is a task that our brain performs very efficiently, but it is still a challenge in computer sciences to model it. Our ability to recognize human actions does not need necessarily a real moving scene as input. We are also able to recognize actions when we watch some point-light stimuli corresponding to joint positions for example. This kind of simplified stimuli was highly used in the phychophysics community in order to obtain a better understanding in the underlying mechanism involved. The neural mechanisms processing *form* or *motion* taking part of biological motion recognition remains unclear. One the one hand, [4] suggests that biological motion can be derived from dynamic *form* information of body postures and without local image motion; On the other hand, [12] proposes a new type of point-light stimulus which suggests that, in this case, only the *motion* information is enough and the detection of specific spatial arrangements of *opponent-motion features* can explain our ability to recognize actions. In fact, in a recent work [13] showed that biological motion recognition can be done with a coarse spatial location of the mid-level optic flow features.

Studies with functional magnetic resonance (fMRI) confirm that biological motion processing does not consider the motion or form information separately. The motion path needs some form feedbacks to perform an accurate categorization. Experiments in [27] showed neural activity uniquely associated with the perception of complex motion as biologically motion. Michels et al. [36], also with fMRI, studied the brain activation related to biological motion. They found that this kind of stimulus with a high form information causes a strong activation of form-processing areas. They measured that the activations in form areas such as FFA/OFA and EBA are dependent on the amount of form information in the input stimuli. Neurophysiological studies also suggest that the biological motion analysis is a combined process between the related dorsal and ventral path in the brain, see [25] for a review.

Visual motion analysis has been studied during many years in several fields as physiology, psychology and computer vision. Many of those studies tried to relate our perception with the activation of the primary visual cortex V1 and extrastiate visual areas as MT/MST. It seems that the area most involved in motion processing is MT, who receives input motion afferents mainly from V1 [22]. Several works such as [16], [18] have established experimentally the spatial-temporal behavior of simple/complex V1 cells and MT cells, in the form of activation maps. With different methods, both have found directionally selective cells sensitive to motion for a certain speed and direction. More properties about MT can be found in the recent survey [8].

Our goal is to show how the information coming from V1-MT neurons (represented by *spike trains*) can be used in order to classify successfully real scenes. Our motivation comes from the idea proposed by several authors such as Thorpe et al [19], [57] who introduced the notion of rank order coding to classify static images. The authors propose that the neural information is coded by the relative order in which these neurons fire. This idea, which was proposed originally for images, can be extended to video streams. The extension to video sequences is not so simple, using rank order coding for each frame is not sufficient. It is

necessary to consider spike trains which include the causality in the temporal information. Looking for a spiking representation for video streams, we present in this paper a simplified spiking model of V1 that allows us to work with sequences of images.

Regarding motion analysis and action recognition, many solution have been proposed in the computer vision community (see [3] for a review). Those approaches often rely on simplified assumptions or parametric models. Some examples include motion body parts tracking [52], [23], motion periodicity analysis [15], [17], [43], [51], event-based analysis [66] and generic human model recovery [26], [30], [47].

A different line of research suggests the analysis of video sequence as space-time intensity volume [5], [67], where the action is characterized by the properties of stacked silhouettes [6], [39], [59]. Our approach does not stack silhouettes along time, but it considers the action as a whole event in time.

Bio-inspired approaches have also been proposed recently to tackle this challenging problem. For example, Giese and Poggio [25] consider general biological models where the brain activity is represented by a continuous scalar variable which is a valid assumption at this level (see e.g. [14]), but does not strictly corresponds to the true neural encoding (which is related to the spike train itself). Following this idea, [53] propose a biological motion recognition system using a neurally plausible memory-trace learning rule. [29] and [45] implemented ICA as a model of simple cells in the visual cortex.

A simplified representation of the visual processing in the brain is to assume the existence of two pathways, the form -*ventral stream*- and the motion pathway -*dorsal stream*-. Both pathways would have a hierarchical structure from low to high processing level. The visual areas involved in the *form* pathway are, e.g., V1, V2, V4. While in the other hand, the main visual areas involved in the *motion* pathway are, e.g., V1, MT, MST.

Under the scope of motion analysis our approach here presented (see Figure 1) is a simplyfied model of the dorsal stream, considering only V1 and MT. The V1 and MT cells as neurons, emit *spikes* when their action potential exceeds the membrane potential (*spiking neurons*). The spikes generated along time for these neurons (*spike trains*) define a neuronal coding giving an estimation of the global activation of the cell according to the input stimuli. Using the spike trains generated by MT cells we propose the construction of a *motion map* coding the motion information of the input stimuli. Then, with the motion maps generated the biological motion recognition is carried out.

Our motivation to use spikes comes from three diffent sources. First, the biological plausibility of the V1-MT neurons. Second, spiking networks allow direct non-linear computations. And third, that not only the mean firing rate can be considered as a measure of the cell activation; higher order measures, e.g., correlation between spike trains can be used as well.

The description of our analog-to-spike V1 model is in Section 2. Section 3 shows our spike-to-spike MT model. The motion representation through motion maps is in Section 4. The validation of our model in the biological motion recognition task is in Section 5. Finally, the discussion and perspectives could be found in Section 6.

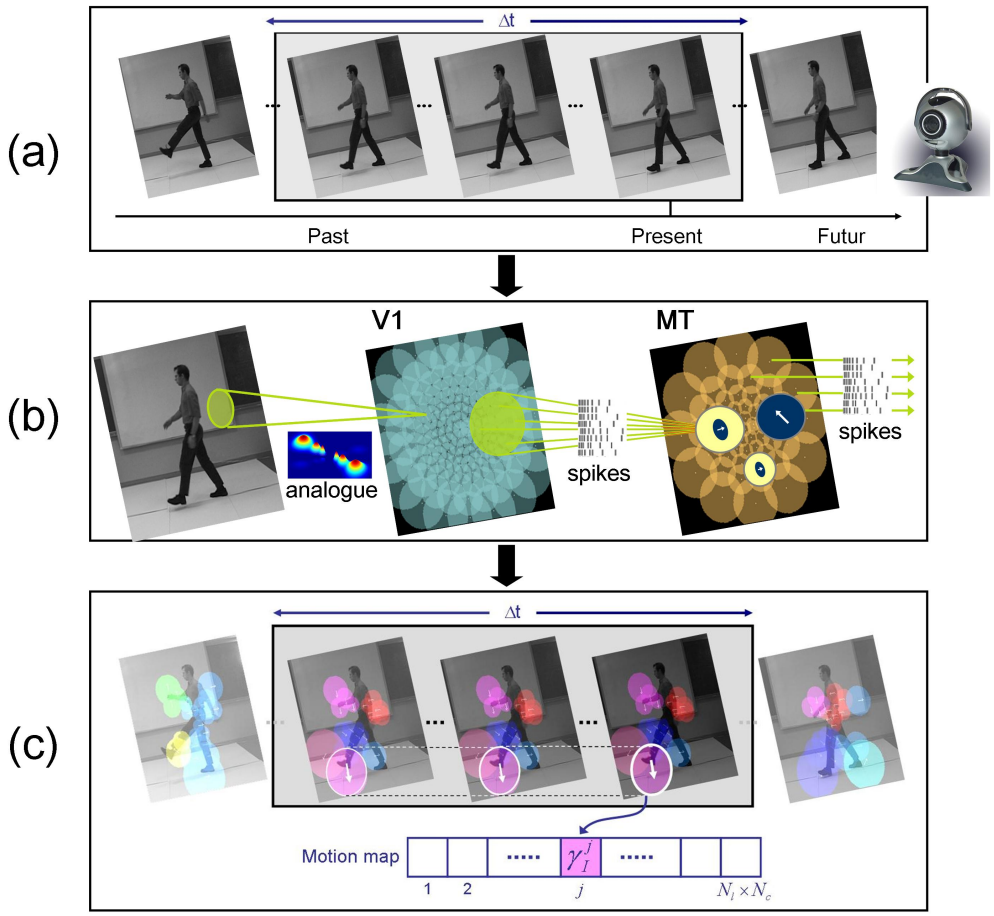

Figure 1: Block diagram showing the different steps of our approach from the input image sequence as stimulus until the *motion map* encoding the pattern motion. (a) We use real video sequence as input, the input sequences are preprocessed in order to have contrast normalization and centered moving stimuli. To compute the *motion map* representing the input image we consider a sliding temporal window of length Δt. (b) Directional-selectivity filters are applied over each frame of the input sequence in a log-polar distribution grid obtaining spike trains as V1 output. These spike trains feed the spiking MT which integrates the information in space and time. (c) The *motion map* is constructed calculating the mean firing rates of MT spike trains inside the sliding temporal window. The *motion map* has a length of $N_L \times N_c$ elements, where N_L is the number of MT layers of cells and N_c is the number of MT cells per layer.

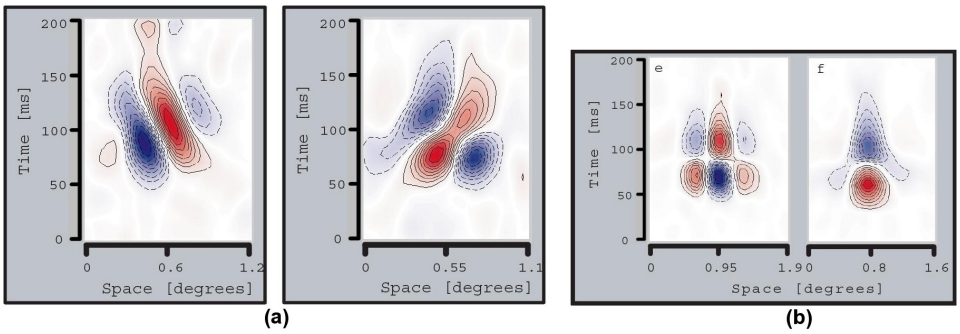

Figure 2: Examples of spatio-temporal receptive fields of V1 cells (from [18]). (a) Spatio-temporal diagrams of two directionally-selective V1 simple cells. (b) Spatio-temporal diagrams of two biphasic V1 simple cells with short latencies.

2 A phenomenological V1 model

2.1 Some facts about V1

Cell measurements in the primary visual cortex show that the extraction of the velocity stimulus is not the first stage of the motion processing in the brain [28]. The motion sensitive cells are directionally-selective and tuned to spatio-temporal frequencies, the velocity (speed + direction) extraction seems to be done in a further stage. In [28] the authors also showed that several properties of simple/complex cells in V1 can be described with energy filters, and in particular using Gabor filters. The individual energy filters are not velocity tuned, however it is possible to use a combination of them in order to have a velocity estimation.

Here, we propose a spiking V1 model built with a bank of energy motion detectors as a local motion estimation. The model is divided in two stages: the analog processing where the motion information is extracted, and the spiking layer where each neuron is modeled as a spiking entity whose inputs are the information obtained in the previous stage. The analog processing is done through energy filters which is a reliable and biologically plausible method for motion information analysis [1]. Each energy motion detector will emulate a complex cell as described in Section 2.2, which will be formed by a non-linear combination of V1 simple cells (see [31] for V1 cells classification). Section 2.4 describes the spiking layer of the model. Finally, the way that all the V1 cells are arranged in order to form V1 is described in Section 2.3, together with the different interactions between V1 cells.

2.2 V1 cells model

2.2.1 Simple cell model

Simple cells are characterized with linear receptive fields where the neuron response corresponds to a weighted linear combination of the input stimulus inside its receptive field.

Combining two simple cells in a linear manner it is possible to get direction-selective neurons, that is, simple cells selective for stimulus orientation and spatial frequency. Since simple cells combine input stimulus using positive a negative weights, the linear receptive fields can have positive or negative responses.

The direction-selectivity (DS) refer to the property of a neuron to respond selectively to the direction of the motion of a stimulus. The way to model this selectivity is obtain receptive fields oriented in space and time. Adding or subtracting neuron responses in spatio-temporal quadrature it is possible to obtain DS simple cells. Let us define the following spatio-temporal oriented simple cells

$$\begin{aligned} F^a_{\theta,f}(x,y,t) &= F^{odd}_\theta(x,y)H_{fast}(t) - F^{even}_\theta(x,y)H_{slow}(t), \\ F^b_{\theta,f}(x,y,t) &= F^{odd}_\theta(x,y)H_{slow}(t) + F^{even}_\theta(x,y)H_{fast}(t), \end{aligned} \tag{1}$$

where simple cells defined in (1) are spatially oriented in an angle θ, and $f = (\bar{\xi}, \bar{\omega})$ is the spatio-temporal orientation in the frequency domain, where $\bar{\xi}$ and $\bar{\omega}$ are the spatio and temporal maximal responses, respectively (see Section 2.2.2). Each conforming simple cell is formed using the first and second derivative of a Gabor function for the spatial part and a Gamma function for the temporal part, respectively, as described below:

$$F^{odd}_\theta(\vec{x}) = \frac{\partial G_\theta(\vec{x})}{\partial x} \tag{2}$$

$$F^{even}_\theta(\vec{x}) = \frac{\partial^2 G_\theta(\vec{x})}{\partial x^2} \tag{3}$$

where $\vec{x} = (x, y)$, $\vec{k} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$, $\omega_f = 2\pi f$, and $G_\theta(x, y)$ corresponds to the Gabor function defined as

$$G_\theta(x,y) = \exp\left(-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}\right)\sin\left(\omega_f\vec{k}\vec{x}\right) \tag{4}$$

where f is the spatial frequency of the gabor function, and θ the spatial orientation.

The temporal contributions $H_{fast}(t)$ and $H_{slow}(t)$ come from the substraction of two Gamma functions with a difference of two in their respective orders.

$$\begin{aligned} H_{fast}(t) &= T_{3,\tau}(t) - T_{5,\tau}(t), \\ H_{slow}(t) &= T_{5,\tau}(t) - T_{7,\tau}(t), \end{aligned} \tag{5}$$

and $T_{\alpha,\tau}(t)$ is defined as

$$T_{\alpha,\tau}(t) = \frac{t^\alpha}{\tau^{\alpha+1}\alpha!}\exp\left(-\frac{t}{\tau}\right), \tag{6}$$

who models the series of synaptic and cellular delays in signal transmission, from retinal photoreceptors to V1 afferents serving as a plausible approximation of biological findings [46].

The biphasic shape of $H_{fast}(t)$ and $H_{slow}(t)$ could be a consequence of the combination of cells of M and P pathways [18], [49] or due to the delayed inhibitions in the retina and LGN [16].

2.2.2 Spatio-temporal frequency analysis of simple cells

Here we analyze the spatio-temporal frequency content of our V1 simple cell model. For simplicity and without loss of generality, we will use just one spatial dimension x and focus on the function $F^a(x)$.

To get the impulse response, we calculate the Fourier transform of F^a, denoted by $\tilde{F}^a$, considering an input stimuli $L(x,t) = \delta(x,t)$

$$\tilde{F}^a(\xi,\omega) = \tilde{F}^{odd}(\xi)\tilde{H}_{fast}(\omega) - \tilde{F}^{even}(\xi)\tilde{H}_{slow}(\omega) \tag{7}$$

Expanding each term we obtain

$$\begin{aligned}
\tilde{F}^{odd}(\xi) &= \frac{\sigma\sqrt{2\pi}}{2}\left\{\exp\left(-\frac{\sigma^2(\xi-\xi_0)^2}{2}\right) + \exp\left(-\frac{\sigma^2(\xi+\xi_0)^2}{2}\right)\right\}, \\
\tilde{F}^{even}(\xi) &= -j\frac{\sigma\sqrt{2\pi}}{2}\left\{\exp\left(-\frac{\sigma^2(\xi-\xi_0)^2}{2}\right) - \exp\left(-\frac{\sigma^2(\xi+\xi_0)^2}{2}\right)\right\}, \\
\tilde{H}_{fast}(\omega) &= \frac{1}{(1+j\omega)^4} - \frac{1}{(1+j\omega)^6}, \\
\tilde{H}_{slow}(\omega) &= \frac{1}{(1+j\omega)^6} - \frac{1}{(1+j\omega)^8},
\end{aligned} \tag{8}$$

To visualize the range of spatial and temporal frequencies of this filter it is necessary to get the power spectrum, which is given by

$$\begin{aligned}
|\tilde{F}^a(\xi,\omega)|^2 &= \frac{\tilde{F}^a(\xi,\omega)\overline{\tilde{F}^a(\xi,\omega)}}{2\pi} \\
&= \frac{1}{2\pi}\left\{|\tilde{F}^{odd}(\xi)\tilde{H}_{fast}(\omega)|^2 + |\tilde{F}^{even}(\xi)\tilde{H}_{slow}(\omega)|^2\right\} \\
&+ \frac{j}{2\pi}\left\{\mathbb{R}\{\tilde{F}^{odd}(\xi)\}\Im\{\tilde{F}^{even}(\xi)\}\left(\tilde{H}_{fast}(\omega)\overline{\tilde{H}_{slow}(\omega)} - \tilde{H}_{slow}(\omega)\overline{\tilde{H}_{fast}(\omega)}\right)\right\}
\end{aligned} \tag{9}$$

The power spectrum (9) is shown in Figure 3. The quotient between the highest temporal frequency activation ($\bar{\omega}$) and the highest spatial frequency ($\bar{\xi}$) corresponds to the speed of the filter. It is also possible to see a small activation for the same speed but in the opposite motion direction. The activation in the anti-preferred direction tuning is an effect also seen in real V1-MT cells data [56], where V1 cells have a weak suppression in anti-preferred direction (30%) compared with MT cells (92%).

As we can see, for a given speed, the filter covers a specified region of the spatio-temporal frequency domain. So, the filter will be able to see the motion for a stimulus whose spatial

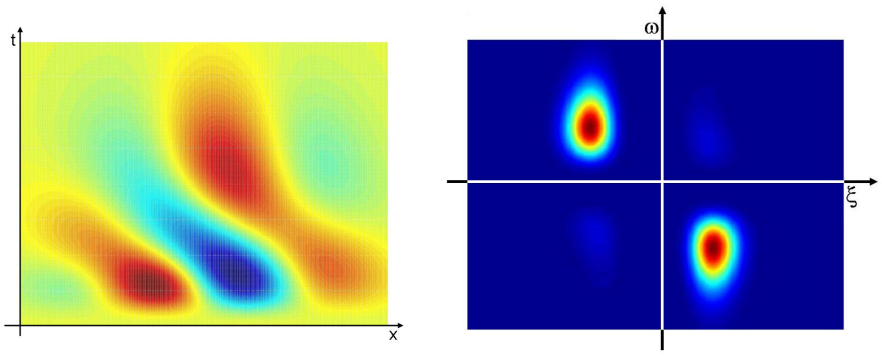

Figure 3: Space-time diagrams for $F^a(x,t)$ (left) and its power spectrum $|\tilde{F}^a(\xi,\omega)|^2$ (right). Both graphs were constructed considering just one spatial dimension x. *Left:* It is possible to see directionality-selective obtained after the linear combination of cells. It is important also to remark the similarities with the biological maps measured by [18] (Figure 2). *Right:* Spatio-temporal energy spectrum of the directional-selective filter $F^a(x,t)$. The slope formed by the peak of the two blobs corresponds to the speed tuning of the filter.

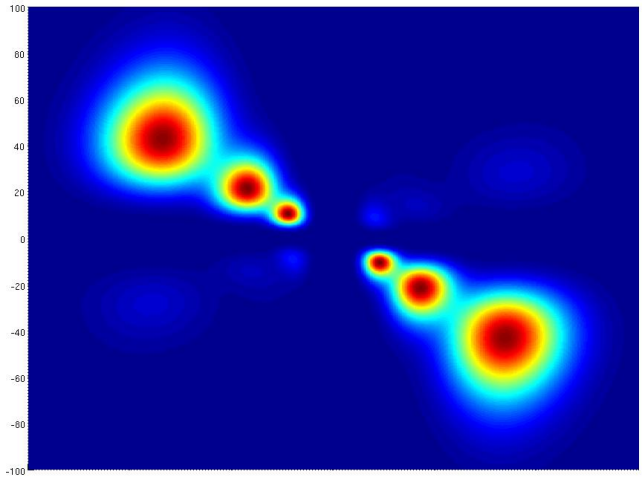

Figure 4: Different filters tuned at the same speed used to tile the spatial-temporal frequency space. This graph was obtained considering just one spatial dimension x.

frequency is inside the energy spectrum of the filter. To pave all the space in a homogeneous way, it is necessary to take more than one filter for the same spatio-temporal frequency orientation. Each filter, for a given orientation, must have different spatial frequencies and thereafter different temporal frequencies to keep the ratio $peak_\omega/peak_\xi$ constant. A diagram with the filter bank tuned at the same speed can be seen in Figure 4.

In our case, the causality of $H_{fast}(t)$ and $H_{slow}(t)$ generates a more realistic model than the one proposed by [54]. Using the temporal components defined in (5), the search of an analytic expression for $|\tilde{F}^a(\xi,\omega)|^2$, is not an easy task, specially due to the non-separability of $F^a(x,t)$.

2.2.3 Complex cell model

Complex cells are also direction-selectivity neurons, however they include other characteristics that cannot be explained by a linear combination of the input stimulus. Their responses are relatively independent of the precise stimulus position inside the receptive field, which suggest a combination of a set of V1 simple cells responses. The complex cells are also invariant to contrast polarity which indicates a kind of rectification of their ON-OFF receptive field responses.

Based on [1], we define our complex cells combining V1 simple cells responses in a nonlinear manner. The combination is done taking the squared sum of a pair of them with the same amplitude response but in spatial quadrature, obtaining an estimation of the local energy of the stimulus. The squared sum of the quadrature filters is independent of the sign of the contrast, it is and constant in time for a drifting sinusoidal as stimulus. A diagram with this procedure can be seen in Figure 5

Let us denote $C_{\mathbf{x}_i,\theta_i,f_i}(t)$ the response of the ith complex cell located at position $\mathbf{x}_i = (x_i, y_i)$, orientation tuning θ_i and spatio-temporal frequencies $f_i = (\bar{\xi}_i, \bar{\omega}_i)$. So, for an input luminosity profile $L(x,y,t)$ the response of the complex cell i, is given by

$$C_{\mathbf{x}_i,\theta_i,f_i}(t) = \left[\left(F^a_{\theta_i,f_i} * L\right)(x_i,y_i,t)\right]^2 + \left[\left(F^b_{\theta_i,f_i} * L\right)(x_i,y_i,t)\right]^2 \tag{10}$$

where the symbol $*$ represents the spatial-temporal convolution, and $F^a_{\theta_i,f_i}$ and $F^b_{\theta_i,f_i}$ are the V1 simple cells defined in (1).

Remark *Velocity Estimation:* As we mentioned in this section, the individual energy filters are not velocity tuned. It is necessary to combine their responses in order to extract the velocity. The spatio-temporal filters defined in (1) can be tuned for a particular orientation in the spatio-temporal domain, but the energy measured is a function of both the velocity and the contrast of the stimulus pattern. In order to do the motion detectors independent of the stimulus contrast, Adelson and Bergen [1] propose a velocity estimation, defined as

$$v = \frac{\sum \left(C_\theta(x,y,t)\right)^2 - \sum \left(C_{\theta+\pi}(x,y,t)\right)^2}{\sum \left(S_\theta(x,y,t)\right)^2} \tag{11}$$

where $S_\theta(x,y,t)$ is the response of a filter tuned for stationary stimulus, same spatial orientation and same spatio-temporal frequencies than $C_\theta(x,y,t)$.

In our model we will not consider the velocity magnitude estimation defined in (11). At V1 level we will just put emphasis in the uniform tile of the frequency space. MT layer will be in charge of the extraction of the velocity stimulus starting from V1 complex cell responses. ■

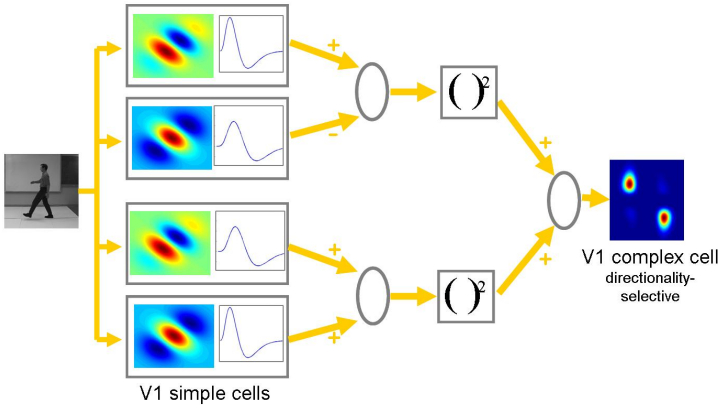

Figure 5: Diagram with the procedure described in [1] to create a V1 complex cell starting from V1 simple cells. At the output a directionally-selective cell is obtained.

2.3 Organization of V1

Given V1 complex cells modeled by (10), we will consider N_L layers of V1 cells. Each V1 layer is modeled as a set of columns of V1 complex cells (see Figure 6):

- Each layer is built with V1 cells with the same spatio-temporal frequency tuning and N_{or} different orientations.
- The related spatio-temporal frequency, and the physical position of the cell inside V1 define its receptive field.
- All the V1 cells belonging to one layer, with receptive fields centered in the position (x_i, y_i), form what we call a *column*.
- One *column* has as many elements as the number of orientations defined N_{or}. For a diagram of the columns of a V1 layer see Figure 6.

The centers of the receptive fields are disposed along a radial log-polar scheme with a foveal uniform zone. The related one-dimensional density $d(r)$, depending of the eccentricity r, is taken as

$$d(r) = \begin{cases} d_0 & \text{if } r \leq R_0, \\ d_0 R_0 / r & \text{if } r > R_0, \end{cases} \tag{12}$$

So that, two regions are defined in (12). The limit between the two regions is given by the value of R_0. The cells with an eccentricity r less than R_0 have an homogeneous density and their receptive fields correspond to the retina fovea (*V1 fovea*). The cells with an eccentricity greater than R_0 have a density depending on r and receptive fields lying outside the retina fovea (*V1 periphery*).

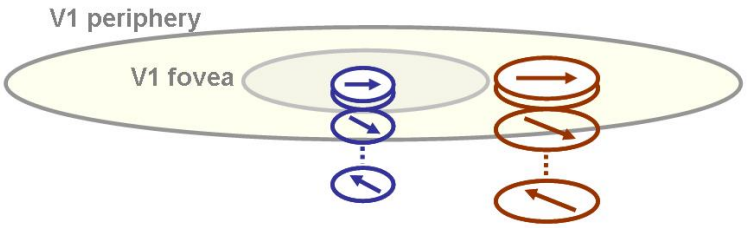

Figure 6: Diagram with the architecture of one V1 layer. There are two different regions in V1, the fovea and periphery. Each element of the V1 layer is a column of N_{or} V1 cells, where N_{or} corresponds to the number of orientations.

2.4 Layer of integrate and fire neurons

The response of the V1 complex cell $C_{\mathbf{x},\theta,f}$ defined in (10), formed as a combination of the V1 simple cells defined in (1), is analogous. To transform the analogous response to a spiking response, the cell will be model as a conductance-driven integrate-and-fire neuron [61], [20].

Considering a spiking V1 complex cell i whose center is located in $\mathbf{x}_i = (x_i, y_i)$ of the visual space, the integrate-and-fire normalized equation is given by

$$\frac{du_i(t)}{dt} = G_i^{exc}(t)(u_i(t) - E^{exc}) + G_i^{inh}(t)(u_i(t) - E^{inh}) - g^L(u_i(t) - E^L) + I_i(t), \quad (13)$$

The neuron i, with orientation tuning θ_i and spatio-temporal frequencies $f_i = (\bar{\xi}_i, \bar{\omega}_i)$, emits a spike when the normalized membrane potential of the cell $u_i(t)$ is equal to 1, then $u_i(t)$ is reinitialized to 0. $I_i(t)$ denotes the external current inputs to the neuron. $G_i^{exc}(t)$ is the normalized excitatory conductance directly associated with the presynaptic neurons connected to neuron i. The conductance g^L is the passive leaks in the cell's membrane. Finally, $G_i^{inh}(t)$ is an inhibitory normalized conductance dependent on, e.g., lateral connections or feedbacks from upper cortical layers. The typical values for the reversal potentials E^{exc}, E^{inh} and E^L are 0mv, -80mv and -70mV, respectively [61].

So, in the particular case of V1, let us consider a spiking V1 complex cell i previously defined. For this neurons, $G_i^{exc}(t)$ is the normalized excitatory conductance directly associated with the pre-synaptic neurons connected to V1 cells. The external input current $I_i(t)$ is here associated with the analogous V1 complex cell response. Finally, $G_i^{inh}(t)$ is an inhibitory normalized conductance dependent on the spikes of neighboring cells of the same V1 layer.

We model the external input current $I_i(t)$ of the ith cell in 13 as the analog response

$$I_i(t) = k_{exc}\Lambda_i(t)C_{\mathbf{x}_i,\theta_i,f_i}(t), \quad (14)$$

where k_{exc} is an amplification factor, $C_{\mathbf{x}_i,\theta_i,f_i}$ refers to the complex cell response defined in (10) and Λ_i summarizes the modulating effect of the neighboring cell interactions provoked mainly by surround suppression interactions [32, 50].

2.5 Surround suppression

The majority of V1 cells present surround modulation, which is normally suppressive [32, 50]. The effect of the surround on the total activation of the neuron can be either subtractive or divisive.

Using the data of 138 cells, Sceniak et al. [50] found that the mean size of the suppression area is of 2.2°. In order to explore the spatial organization of the surround suppression Jones et al. [32] studied the spatial location of suppressive zones in a population of V1 neurons. The majority of the cells analyzed in this way (81%) exhibited spatial heterogenity of surround locations, although 19% showed spatially uniform surround suppression.

Cells exhibiting heterogeneous surrounds, were divided into spatially asymmetric cells (44%) where the surround suppression was biased toward one location and bilaterally symmetric cells (37%) where the surround effect was localized to two opposing regions along a single axis. For cell with heterogenous surrounds, suppressive effects were nearly equally distributed in all directions round the CRF; there was no evidence to suggest that suppressive effects were concentrated in end-zones or side-band regions.

Considering the subtractive or divisive effect of the surround in the activation of a V1 neuron, we will model the surround modulation $\Lambda_i(t)$ of equation (14) as a division of difference of integrated Gaussians. So, considering divisive effect of the surround activation the value of $\Lambda_i(t)$ will be given by

$$\Lambda_i(t) = \left(\frac{k_c}{1 + k_s L_c(t)} \right) \tag{15}$$

where $L_s(t)$ is the surround activation defined as

$$L_s(t) = \int_{-R_s}^{R_s} C_{\mathbf{x}_\psi, \theta_\psi, f_\psi}(t) e^{-\psi^2/2\sigma_s^2} d\psi \tag{16}$$

and where R_s is the radius of the surround suppression area, and σ_s is the corresponding parameter of the Gaussian modeling the surround effect.

3 Introducing a spiking MT model

3.1 Some facts about MT

The middle temporal visual area (MT or V5) of the macaque monkey is an extrastiate visual area in which most cells are selective for the direction of stimulus motion. MT receives input from several areas in the brain [8] mainly from V1 layer, in particular from the layer 4B which is highly directional-selective [37]. MT is retinotopically organized with an emphasis in the fovea, where the half of MT surface is destinated to the processing of the central 15°of the visual field. At a given eccentricity, the MT receptive fields are about 10 times larger than those in V1.

Different kinds of surround geometry of MT receptive fields are observed in the computation of structure of motion. Half of MT neurons have asymmetric receptive fields introducing anisotropies in the processing of the spatial information [33]. The second half of the population examined by [65] has two different symmetries: circular symmetry surround (20% of the population) and bilaterally symmetric surrounds, which correspond to a pair of surrounding regions on opposite sides. The neurons with asymmetric receptive fields seem to be involved in the encoding of important surfaces features, such as slant and tilt or curvature [10].

Regardless the shape of the MT receptive fields, it is possible to classify them according to the interactions between the center and surround [7]. The direction tuning of the surround is broader than that of the center, and the preferred direction, with respect to that of the center, tended to be either in the same (*Reinforcing* surrounds) or opposite (*Antagonistic* surrounds) direction and rarely in orthogonal directions. The antagonistic surrounds are insensitive to wide-field motion but very sensitive to local motion contrast. Otherwise, the reinforcing surrounds have better response to wide-field motion.

MT cells are highly directionally-selective compared with V1. Both V1 and MT have direction tuned neurons, but MT shows a strong inhibition in the anti-preferred direction. The proportion of directionally-selective responses is 30% in V1 and 92% in MT [56].

Compairing the direction selectivity of MT neurons for gratings and plaids, it is possible to classify them as *pattern* direction selective (PDS) or *component* direction selective (CDS). The PDS neurons have a unimodal response for plaids, while the CDS neurons show a bimodal response indicating that two directions of the gratings conforming the plaid stimulus. The fact that the time response of CDS neurons is faster (about 6ms) than PDS neurons (about 50-75ms), suggests a two-stage model for MT, where the outputs of the CDS neurons are used as input of the PDS [38]. The selectivity of a PDS cells evolve during the first 100-150 ms after the exposition of a complex stimulus as plaid [55], starting with a broader selectivity resembling CDS cells. After some tens of milliseconds, their responses evolve to be more PDS-like. By the other hand, CDS cells give a stable response as soon as the stimulus is set.

It is well known that MT cells are tuned to speed, but is this tuning invariant to the spatial frequency of the stimulus? In [44] the authors, over a population of 104 MT cells, found three types of cells: *speed-tuned* neurons, *spatio-temporally independent* neurons and a third group without classification. The *speed-tuned* neurons are motion-sensitive cells invariant to the spatial frequency of the stimulus. The *spatio-temporally independent* neurons are also motion-sensitive cells but tuned for an specific spatial and temporal frequencies, so the speed tuning changes together with the spatial frequency of the stimulus.

For the computation of speed in MT cells, there are two branches of study about how this process is carried out. One group (e.g. [2], [54]) considers that the information coming from V1 cells is linear and MT is who adds the needed non-linearities for the velocity computation. The second group (e.g. [62]) points out that the 2D motion is extracted in V1 through nonlinearities such as endstopping, and MT just polls this information. In addition, there is evidence that the overall level of speed is modulated by the surround motion [7].

The energy model used for V1 simple/complex cells representation does not allow them to be sensitive to an specific speed. In space they are sensitive just to the orthogonal component of their preferred spatial orientation. A velocity-selective neuron may be constructed polling the output of several V1 complex cells with the spatio-temporal orientation consistent with the velocity desired.

3.2 Spiking MT model

3.2.1 Connections to V1 cells

Our model is a spiking neural network where each entity or node is a MT cell. Each MT cell i can be modeled as conductance-driven integrate-and-fire neuron similar to (13).

$$\frac{du_i(t)}{dt} = G_i^{exc}(t)\,(u_i(t) - E^{exc}) + G_i^{inh}(t)\left(u_i(t) - E^{inh}\right) + g^L\left(u_i(t) - E^L(t)\right) + I(t) \quad (17)$$

where $u_i(t)$ is the normalized membrane potential of the cell. G_i^{exc} refers to the normalized excitatory conductance associated to an excitatory reversal potential E^{exc}. Similarly, G_i^{inh} is the normalized inhibitory conductance associated to an inhibitory reversal potential E^{inh}. g^L denotes the passive leaks in the cell's membrane associated to the reversal potential $E^L(t)$. $I(t)$ denotes the external current inputs to the neuron. Like in (13) there is also a reinitialization of the neuron membrane potential to zero as soon as its voltage reaches the threshold, which in this case for normalizations effects the threshold is considered as 1. Typical values for the conductances E^{exc}, E^{inh} and E^L are 0mv, -80mv and -70mV, respectively [61].

The neuron i is a part of a spiking neural network where the input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ are obtained considering the activity of all the pre-synaptic neurons connected to it. For example, if a pre-synaptic neuron j has fired a spike at time $t_j^{(f)}$, this spike reflects an input conductance to the post-synaptic neuron i with a time course $\alpha(t - t_j^{(f)})$. In our case the pre-synaptic neurons corresponds to the V1 outputs (see Figure 8(a)). According to this, the total input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ of the post-synaptic neuron i are expressed as

$$\begin{aligned} G_i^{exc}(t) &= \sum_j w_{ij}^+ \sum_f \alpha(t - t_j^{(f)}) \\ G_i^{inh}(t) &= \sum_j w_{ij}^- \sum_f \alpha(t - t_j^{(f)}) \end{aligned} \quad (18)$$

where the factor w_{ij}^+ (w_{ij}^-) is the efficacy of the positive (negative) synapse from neuron j to neuron i (See [24] for more details). The time course $\alpha(s)$ of the postsynaptic current in (18) can be modeled as an exponential decay with time constant τ_s as follows

$$\alpha(s;\tau_s) = \left(\frac{s}{\tau_s}\right) exp\left(-\frac{s}{\tau_s}\right) \quad (19)$$

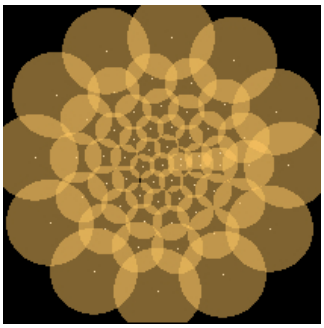

Figure 7: Sample of log-polar architecture used for a MT layer. The cell distribution law is divided into two zones, a homogeneous distribution in the center with a certain radius and then a periphery where the density of cells decays with the eccentricity.

Inserting (18) into (17) we finally obtain

$$\frac{du_i(t)}{dt} = \left\{ \sum_j w_{ij}^+ \sum_f \alpha(t - t_j^{(f)}) \right\} (u_i(t) - E^{exc}) + \left\{ \sum_j w_{ij}^- \sum_f \alpha(t - t_j^{(f)}) \right\} \left(u_i(t) - E^{inh}\right) + g^L \left(u_i(t) - E^L(t)\right) + I(t) \quad (20)$$

Each MT cell has a receptive field from where converge V1 complex cells afferents inside its receptive field, which correspond to the pre-synaptic neurons j in (20). Those inputs will be excitatory ($w_{ij}^+ > 0$) or inhibitory ($w_{ij}^- < 0$) depending on the characteristic and shape of the respective receptive fields [65, 63].

The receptive field associated with an MT cell corresponds to a certain area inside the visual field. The half of MT surface is assigned to process the information coming from the central 15° of the visual field, which receptive field size of a MT cell inside this region is about 4-6 times bigger than the V1 receptive field [35].

The MT cells are distributed in a log-polar architecture, with a homogeneous area of cells in the center and a periphery where the density decreases with the distance to the center of focus. While the density of cells decreases with the eccentricity, the size of the receptive fields increases preserving its original shape. Figure 7 shows an example of the log-polar distribution of MT cells.

Different layers of MT cells conform our model. Each layer is built with MT cells of the same characteristics, this is same speed and direction tuning. Depending on the tuning values, the MT cell decides which V1 cells contribute with relevant information and establishes the proper connection between them. The criteria of selection is to consider all the V1 cells inside the MT receptive field with an absolute difference of motion direction-selectivity respect to MT cell no more than $\pi/2$ radians. The weight associated to the connection between neuron j and i is proportional to the angle α_{ij} between the two preferred motion direction-selectivity (see Figure 8(b)). The connection weight w_{ij} between the jth V1 cell

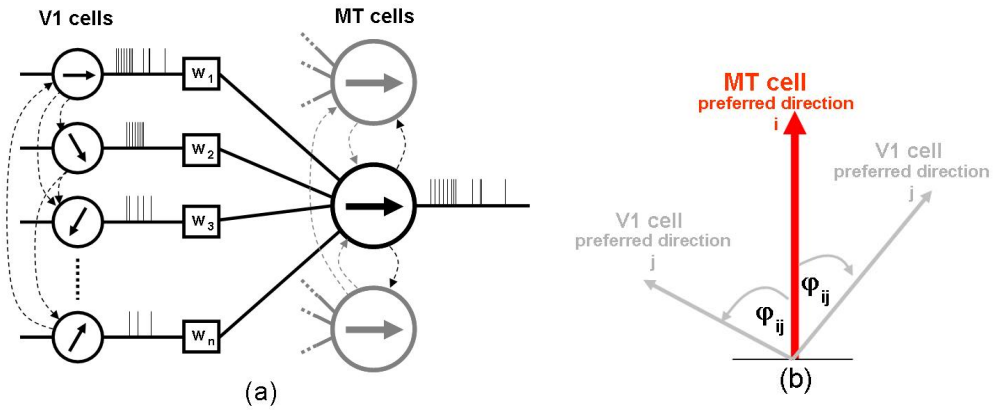

Figure 8: Architecture of the spiking neural network to model MT. Each MT cell receives as input the afferent V1 cells (a). The connection weights between V1 and MT cells are modulated by the cosine of the angle α between the preferred direction of MT cell and the preferred direction of V1 cell (b).

and the ith MT cell is given by

$$w_{ij} = \begin{cases} k_c w_{cs}(\mathbf{x}_i - \mathbf{x}_j)\cos(\alpha_{ij}) & \text{if } 0 \leq \alpha_{ij} \leq \frac{\pi}{2}, \\ 0 & \text{if } \frac{\pi}{2} \leq \alpha_{ij} \leq \pi, \end{cases} \tag{21}$$

and where k_c is an amplification factor, α_{ij} corresponds to the absolute angle between the preferred jth MT cell direction and the preferred ith V1 cell direction. $w_{cs}(\cdot)$ is the weight associated to the difference between the center of MT receptive field $\mathbf{x}_i = (x_i, y_i)$ and the V1 cell center position $\mathbf{x}_j = (x_j, y_j)$. The value of $w_{cs}(\cdot)$ depends on the shape of the receptive field associated to the MT cell (see Section 3.2.3). Depending on the sign of $w_{cs}(\cdot)$ we define

$$\begin{aligned} w_{ij}^{+} &= w_{ij}, \text{ if } w_{cs} > 0 \\ w_{ij}^{-} &= w_{ij}, \text{ if } w_{cs} < 0 \end{aligned} \tag{22}$$

Remark Concerning conectivity within and between V1-MT cells, [42] propose the inclusion of an intermediate *V1 cluster* between V1 and MT cells. The *V1 cluster* is created in order to justify the behavior of MT pattern/component neurons reported in [34]. The *V1 cluster* recolects V1 neurons both motion reinforcing and motion opponents units. The MT neuron concentrate several *V1 cluster* entities inside its receptive field. The property response of the MT cell (component/pattern) will depend on the spatial location of the input stimulus inside its receptive field. It is important to remark the role of the motion opponents connections in the *V1 clusters*, these connections block the component response in the plaid case but not in the pseudoplaid case (see [42] and [34] for details). ■

3.2.2 Horizontal connectivity

Diffusion is a biological mechanisms through the cells transmit some distinctive and important information to the neighboring ones. Diffusion also occurs between neighboring cells with different characteristics like different velocities, speed, receptive field. Diffusion of activity between neighboring cells is observed and it is due to a local horizontal connectivity patter. The diffusion can be included in the dynamic of the neuron extending (20) as follows

$$\begin{aligned}\frac{du_i(t)}{dt} &= g^L\left(u_i(t) - E^L(t)\right) + I(t) + \left\{\sum_{j\in \mathrm{V1}} w_{ij}^{+} \sum_f \alpha(t - t_j^{(f)})\right\}(u_i(t) - E^{exc}) \\ &+ \left\{\sum_{j\in \mathrm{V1}} w_{ij}^{-} \sum_f \alpha(t - t_j^{(f)})\right\}\left(u_i(t) - E^{inh}\right) + \left\{\sum_{k\in \mathrm{MT}} \zeta_{ik}^{+} \sum_f \alpha(t - t_k^{(f)})\right\}(u_i(t) - E^{exc}) \\ &+ \left\{\sum_{k\in \mathrm{MT}} \zeta_{ik}^{-} \sum_f \alpha(t - t_k^{(f)})\right\}\left(u_i(t) - E^{inh}\right)\end{aligned} \tag{23}$$

where the two last terms correspond to diffusion. ζ_{ik}^{+} (ζ_{ik}^{-}) corresponds to a weight matrix with the positive (negative) diffusion shape modeled.

For the implementation of diffusion it is necessary create connections between MT cells. The connection radius and weights are given by the diffusion function, which is normally a Gaussian. As a first step inside our model the MT cells connected for diffusion are the cells with the same velocity direction. So, inside a neighborhood all the cells with the same velocity direction share information trough diffusion without concerning the different speeds of the cells.

The information exchange between MT cells is done through spikes, where the first cell who fires send the diffusion signal to the assigned neighbors.

3.2.3 Receptive fields: geometry and dynamics

The geometry and dynamics of the MT receptive fields is far from be completely understood. Their geometry is the main responsible of the direction tuning of the MT cell and it changes along time, either switching from component to pattern behavior [55] or showing a direction reversal from preferred to antipreferred direction tuning [41].

The contrast of the input stimulus plays an important role in the MT velocity tuning. Experiments done by [40] showed that the velocity tuning of the MT cell is modulated by the contrast of the input stimulus. They found two types of cells: The first group of cells has to low input a broad tuning to low speed switching to high speed tuning when the input contrast is high. The second group has the opposite behavior, this means at low contrast the cells are broadly tuned to high speeds while at high contrast they are highly tuned to low speed.

A second level of complexity in the MT receptive fields is their spatial organization. Experiments in [65] revealed that only 20% of the MT cells tested have circularly symmetric surrounds, 50% were asymmetric concentrating the suppression in only one location on one side of the preferred-null direction axis, and 25% have bilaterally symmetric zones of suppression lying along the same axis (see Figure 9). In our case we will just consider circularly symmetric surrounds. Altough this kind of surrounds are just the 20% of the MT cells tested by [65], it looks to be a good inital approximation for the motion recognition task (see Section 5).

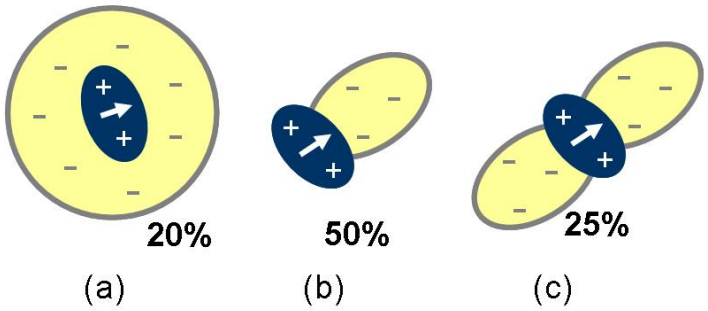

Figure 9: Different geometries of asymmetric center-surround organization in MT cells [65, 64]. (a) Circularly symmetric surrounds. (b) Asymmetric configuration concentrating the suppression at one side of the motion preferred axis. (c) Bilaterally symmetric zones of suppression lying in the motion preferred axis.

Regarding organization and center-surround interactions, [7] shows two different types of cells, the reinforcing surrounds and the antagonistic surround. The direction tuning of the surround is always broader than the center. The direction tuning of the surround compared with the center tends to be either the same or opposite, but rarely orthogonal. The antagonistic surrounds are insensitive to wide-field motion but sensitive to local motion contrast. By the other hand, the cells with reinforcing surround are best sensitive to wide-field motion. The author in [7] also suggests a columnar organization in MT cells, grouping the columns according to its center-surround properties.

Considering the results found by [7], we propose three types of MT center-surround interactions in our model. Our claim is that the antagonistic surrounds contain key information about the motion characterization, which could highly helps the motion recognition task. We propose one reinforcing center-surround interaction and two antagonistic as shown in Figure 10. It is important to mention that this approach corresponds to a coarse approximation of the real receptive field shapes, but anyways this approach is capable to extract key information from the motion stimulus. The receptive fields shown in Figure 10(a) and Figure 10(b)-(c) were created using a Gaussian and a Difference of Gaussians (DoG), respectively.

We put at the entrance of our system a stimulus consisting of a circle of variable radius with a drifting grating inside. The motion direction of the drifting grating corresponds to motion direction tuning of our MT cells. We varied the radius of the circle and we measured

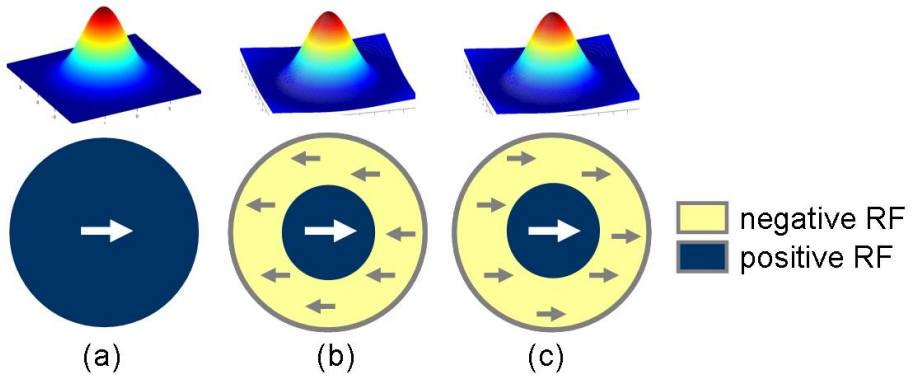

Figure 10: Center-surround interactions modeled in the MT cells. The reinforcing surround (a) is modeled through a gaussian. The two receptive fields with inhibitory surround (b), (c) are modeled with a Difference-of-Gaussians. The cells with inhibitory surround have either antagonistic direction tuning between the center and surround or the same direction tuning.

the firing rate of the cell. The graphs with the firing rate measured according to the radius of the stimulus can be seen in Figure 11. A more realistic receptive fields shapes can be obtained using also difference-of-gaussian and trying to fit them with the real data collected by [9], [7], [65], or [64].

4 Motion maps

We proposed in previous sections a bio-inspired MT model where a video stream is converted into a set of spike trains. Our claim is that the information contained in those spike trains corresponds to a discrete representation of the motion information contained in the input sequence. The idea of coding through spike trains has been previously proposed by Thorpe et al ([57], [48], [**?**]), who introduced the notion of rank order coding to classify static images. The authors claim that the neural information is coded by the relative order in which these neurons fire. The direct extension to video sequences is not so simple, since the use of rank order coding cannot be easily applied to each frame due to time overlapping. It is necessary to consider spike trains which include the causality in the temporal information. Here, we propose a novel representation which summarizes the motion characteristics of a given sequence based in these spike trains.

A motion representation starting from spike trains is not an easy challenge: The difference of phase between two spike trains, the time of the first spike emitted, the mean firing rate (spikes/time window) of a neuron or the synchrony between spike trains of neuron populations could be examples of different information coding in the brain. In our case, it is required to find which MT cells responses are the most representative of a certain motion sequence. Visualizing the video stream as a continuous in time, we could define for an input

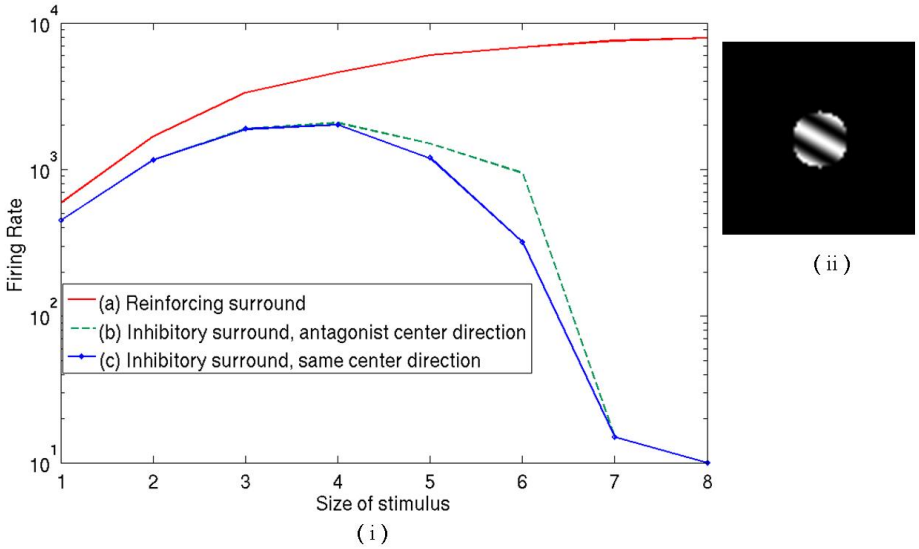

Figure 11: Graph with the firing rates measured for the three different receptive fields modeled for MT cells (Figure 10). The stimulus consisted in a circle of variable radius with a drifting grating inside (ii). The effect of vary the radius of the stimulus in the firing rate of the cell is graphed in (i). Each curve corresponds to the receptive fields (a), (b) and (c) defined in Figure 10. The motion direction tuning of the center of each MT cell corresponds to the orthogonal direction of the drifting grating. Here it is possible to see the no-reaction for wide-field motion in the case of antagonistic surrounds cells.

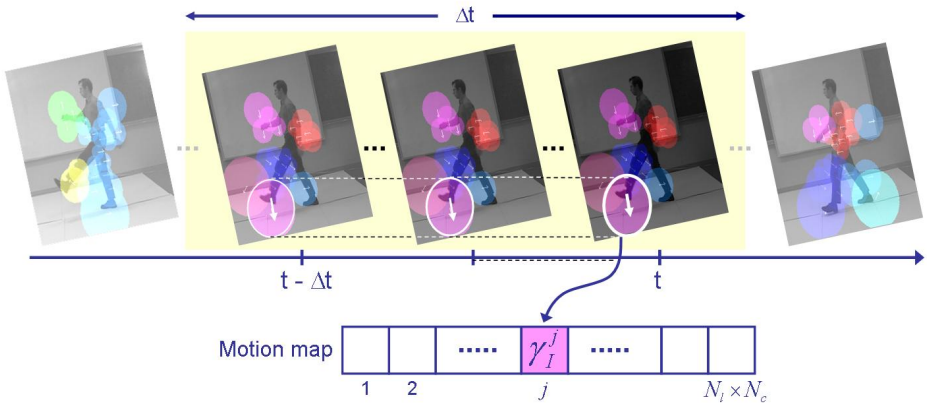

Figure 12: Diagram summarizing the construction of the motion map H_I. A sliding window of width $\triangle t$ is defined. Inside that window, for the j-th cell the mean firing rate $\gamma_I^j(t)$ is calculated. Finally, the motion map is built using the windowed firing rate $\bar{\gamma}_I^j(t)$ of each j-th cell.

sequence I and per each MT cell j, the *windowed firing rate* γ_I^j as

$$\gamma_I^j(t, \triangle t) = \frac{1}{\delta t \triangle t} \sum_{i=t-\triangle t}^{t} \delta(t - t_j^{(i)}), \tag{24}$$

where $\triangle t$ is the sliding window size, δt is the time between frames and $\delta(\cdot)$ is the Kroenecker delta function representing the ith spike emitted by the jth MT cell.

So, following the idea proposed in [21], we define the *motion map* representing the input stimuli $I(x, y, t)$ as

$$H_I(t, \triangle t) = \left\{ \gamma_I^j(t, \triangle t) \right\}_{j=1,\ldots,N_l \times N_c}, \tag{25}$$

where N_l is the number of MT layers and N_c is the number of MT cells per layer. Each element γ_I^j with $j = 1, ..., N_l \times N_c$ is the corresponding windowed firing rate defined in (24). The H_I definition can be summarized in Figure 12.

The task where we want to focus our attention is in biological motion recognition. To do this, we need to classify the *motion maps* obtained in order to create the different classes. To group our samples in classes and do a further recognition, we use a measure discrimination to evaluate the similarities between two motion maps. The comparison between two motion maps $H_I(t, \triangle t)$ and $H_J(t', \triangle t')$, is done using the following expression

$$\mathcal{D}\left(H_I(t, \triangle t), H_J(t', \triangle t')\right) = \frac{1}{N_l \times N_c} \sum_{l=1}^{N_l \times N_c} \frac{\left(\gamma_I^l(t, \triangle t) - \gamma_J^l(t', \triangle t')\right)^2}{\gamma_I^l(t, \triangle t) + \gamma_J^l(t', \triangle t')}. \tag{26}$$

This measure is defined as the *triangular discrimination* introduced by [58].

Another measures derivated from statistics, such as *Kullback and Leiber* (KL) would also be used. The experiments done using the KL measure showed no significant improvements.

Remark The representation shown in (25) is invariant to the sequence length an its starting point. It is also included information regarding the temporal evolution of the activation of MT cells, respecting the causality in the order of events. The fact of use a sliding window allows us to include motion changes inside the sequence. ■

5 Experiments

The architecture proposed for the whole model, starting from the input sequence and ending with the motion map representation H_I, is represented in Figure 1. The system receives as input a sequences of images where the biological motion is carried out. The directional-selective V1 filters are applied over each frame of the sequence in a log-polar distribution grid, using one layer per each velocity. The spike trains generated feed the MT layers where each MT cell is activated according to the activation in V1 stage. The MT cells are arranged in a log-polar grid as well, working joint with V1 cells as a spiking network. Per each input frame the firing rate of all MT cells is calculated, then the mean firing rate along the whole sequence is obtained in order to construct the motion map described in (25). This motion map characterizes and codes the biological motion stimulus.

We ran the experiment using two databases: Giese and Weizmann. The V1 and MT parameters change depending on the database. For *Giese* database we consider 20 different samples of the same subject walking and another 20 samples of marching. Each sequence contains 30 frames. The size of the images that form the sequence is 210x210 pixels. *Weizmann* database is formed by 9 different samples of different persons doing 10 different actions as: bend, jack, jump, pjump, run, side, skip, walk, wave1 and wave2. A representative frame of each action can be seen in Figure 13. Each sequence contains at least 18 frames and the original video streams were resized to have also images of 210x210 pixels.

The experiment protocol is equal to all the tests done. Each experiment considered randomly training sets with the same number of elements per class. The remaining sequences were used to construct the test set. For each number of elements in the training set, the experiment was repeated 20 times obtaining the mean error recognition, the best recognition rate and the standard deviation. All the motion maps of the training set were obtained and stored in a data container. When a new input sequence belonging to the test set is present to the system, the motion map is calculated and it is compared using (26) with all the motion maps stored in the training set. The input sequence class will be selected as the same class of the sequence with the smallest distance (see (26)). This selection mechanism corresponds to a RAW classifier.

Remark We repeated the experiments using a different classifier as SVM, but we did not get significant

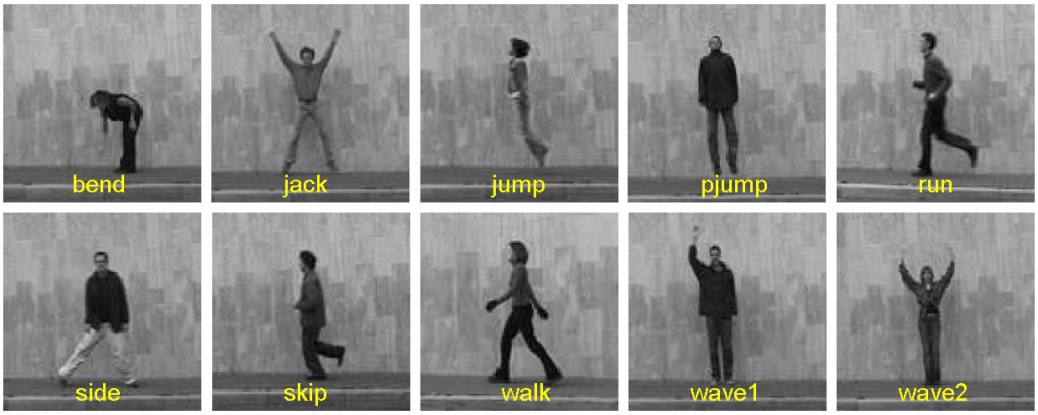

Figure 13: Sample frames of each of the ten actions conforming the Weizmann database (`http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`).

improvements in the recognition performance. ■

5.1 Ground Truth of the Model

The ground truth of our approach was created running the architecture described in Figure 1. In this first simulation we did not consider any inhibition interaction at V1 level, and neither any diffusion in MT level or feedback from MT to V1. The straight-forward procedure ran in order to obtain the basic error rate results.

The settings for V1 stage are the same both for Giese and Weizmann database. For both databases 9 layers of V1 cells were used, each layer with 8 orientations giving a total of 26416 cells per layer (3302x8). The radius of the V1 fovea R_0 (12) was set as 80 pixels and a total V1 size of 200 pixels of diameter. Following the biological fact mentioned in [60] the value of σ corresponds to $1.324/(4\pi f)$. The specific parameters for each V1 layer are listed in Table 1. The settings for MT are listed in Table 2.

Varying the number of samples in the training set, we ran our system and we obtained for Giese and Weizmann databases the recognition error rates shown in Figures 14, 15, respectively.

In order to see the influence of the information coded by more complex receptive field organizations, we repeated the experiments using the three receptive fields shown in Figure 10. The results obtained for this experiments can be seen for Giese and Weizmann databases, in Figure 16 and Figure 17, respectively.

As it is possible to see in Figures 16 and 17, the improvement obtained with the addition of the more complex receptive fields is different for Giese and Weizmann databases. The

V1 settings

	d_0	σ	τ	f	k_{amp}
Layer 1	0.4	0.3323	0.0080	0.3170	8
Layer 2	0.4	0.6647	0.0160	0.1585	8
Layer 3	0.4	1.3295	0.0333	0.0816	8
Layer 4	0.4	0.4214	0.0051	0.2050	8
Layer 5	0.4	0.8429	0.0103	0.1025	8
Layer 6	0.4	1.6857	0.0215	0.0536	8
Layer 7	0.4	1.0250	0.0045	0.1028	8
Layer 8	0.4	2.0498	0.0094	0.0514	8
Layer 9	0.4	4.0996	0.0175	0.0303	8

Table 1: Configuration values used per each V1 layer. The values corespond to those ones defined in (2), (12), (**??**) and (6).

MT settings

	Giese	Weizmann
Radius fovea	20[pixels]	40[pixels]
Cell density in fovea	0.08[cells/pixel]	0.1[cells/pixel]
Eccentricity decay	0.02	0.02
Radius receptive field in fovea	9[pixels]	9[pixels]
Number orientations	8	8
Number cells per layer	60	161

Table 2: Settings used in MT to run the ground truth recognition error rates.

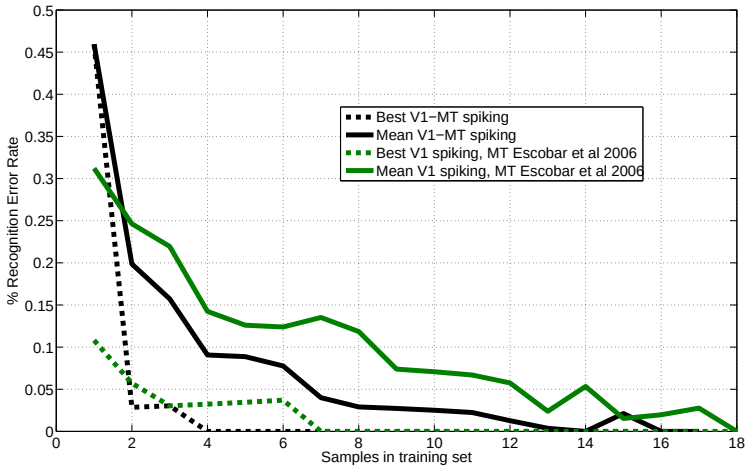

Figure 14: Recognition error rate obtained for the ground truth configuration and the Giese database. In this case just one receptive field (*reinforcing*) were used. It is also included the comparison with the results obtained in [21].

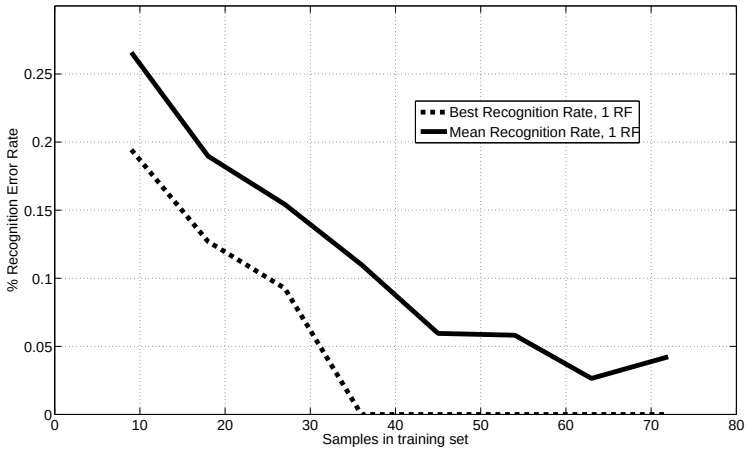

Figure 15: Recognition error rate obtained for the ground truth configuration and the Weizmann database. In this case just one receptive field (*reinforcing*) were used.

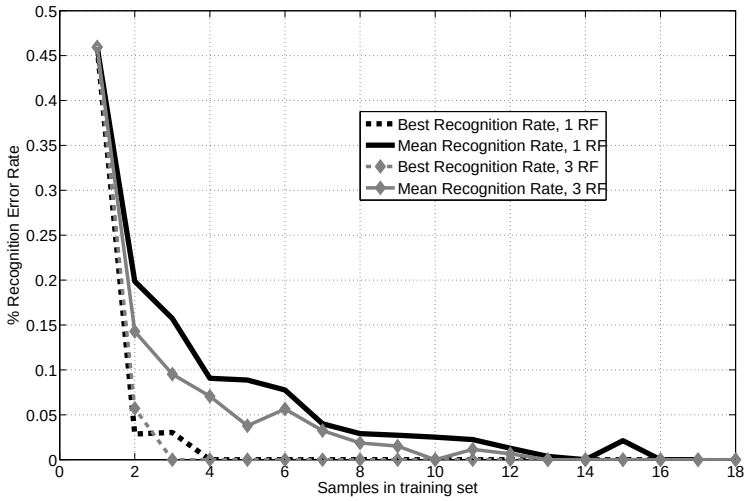

Figure 16: Recognition error rate obtained for Giese database using the three different receptive fields described in Figure 10. It is possible to see an improvement in the recognition error rates in comparison with the case of just one receptive field (Figure 14).

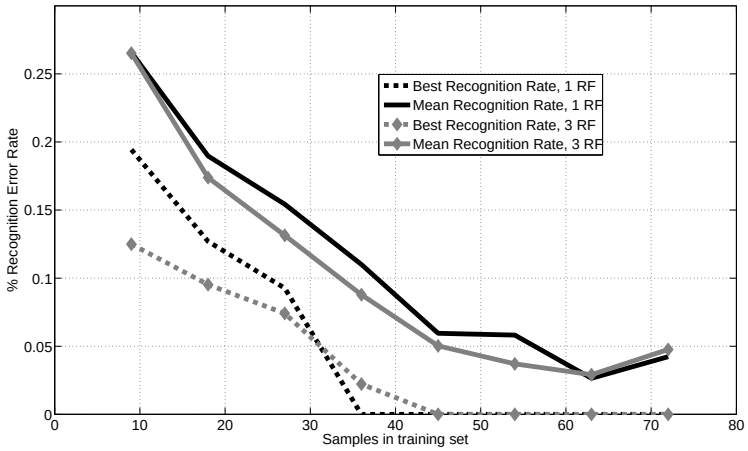

Figure 17: Recognition error rate obtained for Weizmann database using the three different receptive fields described in Figure 10. The improvement obtained with the approach of the three receptive fields instead only one is considerably.

fact that Giese database only contains two classes, gives a good recognition performance in the case of only one receptive field. The recognition performance is improved if we add the more complex receptive fields, but the improvement is not very notoriously. In the case of Weizmann database, which contains ten different classes, we obtained an evident improvement adding the new receptive fields.

6 Discussion and perspectives

We have proposed a spiking V1-MT model. The model receives as input a sequences of images. The V1 cells activated generate spike train feeding the next MT spiking layer. According to the activation of the MT cell, we proposed a motion representation (*motion map*) encoding the information needed for a motion categorization task.

Our spiking V1 model is built with a bank of energy motion detectors as local motion estimators. The V1 model is divided in two stages: the analog processing where the motion information is extracted, and the spiking layer where each neuron is modeled as a spiking entity whose inputs are the information obtained in the previous stage.

The local motion estimation is done through the combination of different spatio-temporal filters. The spatio-temporal filters are combined in order to obtain directional-selectivity (DS) properties. The DS refer to the property of a neuron to respond selectively to the direction of the motion of a stimulus. This property can be obtained combining different spatio-temporal filters. The construction of the spatio-temporal filters, and further DS ones, is inspired by [1]. The spatial parts of the filters are modeled by Gaussian derivatives. The temporal part is characterized by a biphasic shape, which it could be a consequence of the

combination of cells of M and P pathways [18], [49] or due to the delayed inhibitions in the retina and LGN [16].

The V1 spiking layer receives as input the filter response of each V1 directional-selectivity cell. Each spiking cell is modeled as a leaky-integrate-and-fire neuron whose output feed the upper spiking MT layer. There are connections between the different V1 cells, specially to create inhibitory interactions as cross-orientation inhibition [11].

Different layers of V1 cells are used, each of them with a different spatio-temporal frequency tuning. The spatio-temporal frequency distribution of each layer is done in order to tile the whole frequency space of interest. Cells with the same spatio-temporal badwiths and different spatial orientations are considered to be part of a *column*. The V1 columns are arranged in a radial log-polar scheme paving the visual field.

Regarding MT, the model proposed is a spiking neural network where each node corresponds to a MT neuron. The nature of the MT cell, its receptive field geometry and its center-surround interactions define the subset of V1 cells to be connected and their respective connection weights.

In connectivity, each MT neuron is connected with a subset of V1 neurons inside its receptive field, whose spike trains outputs are the input of the MT cell. MT neurons are also connected between them, allowing interactions such as spatial diffusion. The information is propagated within a radius of action following a Gaussian law.

Each MT neuron has a receptive field geometry and a center-surround interaction coding a certain motion information. The shape and interaction of the MT receptive fields are chosen considering the results obtained by [7].

Considering the spike trains generated by the MT cells we propose a motion map as a representation of the motion information contained in the input stimulus. The motion maps summarize along time the activation of the different MT cells. We claim that our motion map represents the input stimuli and can be used in a motion categorization task. These motion maps are invariant to the input sequence lengths and their starting point.

In order to validate our motion map as a valid motion representation, we applied the model to biological motion recognition. We tested the model with two different databases (Giese and Weizmann), obtaining the results shown in Section 5. The good recognition performance obtained with our spike-to-spike model reinforces our hypothesis about the representability of our motion maps. It is also possible to see that the motion information coded by the different receptive field shapes and interactions of MT cells is a key issue in motion categorization.

Giese database, with only two classes, has a good recognition performance nevertheless the center-surround interactions of the MT cells. Of course, the inclusion of more complex interactions improve the recognition performance, but the improvement is not significant compared to the case of the reinforcing center-surround interaction only (see Figure 10). Weizmann database, formed by ten different classes (Figure 13), is a different scenario. The inclusion of center-surround interactions coding more complex motion patterns is a key issue in the recognition performance. The complexity of the variety of classes suggests the use of a more complex model to obtain motion maps more representatives of the input stimulus.

The results here described were obtained using no interactions within V1 or MT cells. These interactions will be tested in order to validate our model with real cell measurements. For this, we will use standard input stimuli such as drifting gratings, plaids or barber poles. Also the inclusion of more complex receptive field geometries in MT cells will be implemented and tested.

Some dynamics in the behavior of MT cells will be considered. Information contained in the neurons described by [41] could code important events for the motion categorization and further recognition. The facts that pattern selectivity cells could come from the activation of the component selectivity cells, and that the pattern cells receive inputs from the ventral stream areas as V2 and V3 [55], give us an idea of the different connections and interactions between different areas of the visual cortex. The changes in MT cell responses to different input contrast [40] could be an important issue for robustness.

Acknowledgments

The dataset has been kindly provided by Dr. Giese and the present work has been realized thanks to this data set. This work was partially supported by the EC IP project FP6-015879, FACETS and CONICYT Chile. We also thank to Olivier Rochel for his Mvaspike simulator, this tools allowed us to create and simulate spiking networks in an easy manner.

Appendix

A Spatio-temporal filters frequency analysis

As we previously mentioned in Section 2.2.3, the directionally-selective complex cell comes from the relation (10)

$$C_\theta(x,y,t) = \left[F^a_\theta(x,y,t) * I(x,y,t)\right]^2 + \left[F^b_\theta(x,y,t) * I(x,y,t)\right]^2,$$

where the spatial and temporal bandwiths are given by the frequence response of $F^a_\theta(x,y,t)$ and $F^b_\theta(x,y,t)$, both not separable. An alternative analysis to the one in Section 2.2.2 could be done through complex analysis. In order to look for separability, let us study the frequence response of the following complex number

$$g_\theta(x,y,t) = F^a_\theta(x,y,t) + jF^b_\theta(x,y,t), \tag{27}$$

which is separable in space and time, it means

$$g_\theta(x,y,t) = g^x_\theta(x) g^y_\theta(y) g^t(t) \tag{28}$$

Considering $\theta = 0$ we can write $g^x_\theta(x)$, $g^y_\theta(y)$ and $g^t(t)$ as follows

$$\begin{aligned}
g^x(x) &= -\frac{\exp\left(-\frac{x^2}{2\sigma^2}\right)}{\sigma^4}\sin(2\pi f x)\left[-x^2+\sigma^2+4\pi^2 f^2\sigma^4 - j\sigma^2 x\right] \\
&\quad - \frac{\exp\left(-\frac{x^2}{2\sigma^2}\right)}{\sigma^4} 2\pi f\sigma^2 \cos(2\pi f x)\left[2x + j\sigma^2\right] \\
g^y(y) &= \exp\left(-\frac{y^2}{2\sigma^2}\right), \\
g^t(t) &= \frac{1}{5040\tau^8}^3 \exp\left(-\frac{t}{\tau}\right)\left[840j\tau^4 + (-42-42j)t^2\tau^2 + t^4\right]\Theta(t),
\end{aligned} \tag{29}$$

The separability property of $g_\theta(x,y,t)$ allows us to study its frequence response considering the frequence response of each of its components g^x_θ, g^y_θ and g^t, separately. For an easier analysis, and without loss of generality, we will consider just one spatial component $g^x(x)$ with $\theta = 0$. Applying the Fourier transform to $g(x,t)$ we get $\tilde{g}(\xi,\omega)$ defined as

$$\tilde{g}(\xi,\omega) = \tilde{g}^x(\xi)\tilde{g}^t(\omega), \tag{30}$$

where $\tilde{g}^x(\xi)$ and $\tilde{g}^t(\omega)$ are the Fourier transforms of $g^x(x)$ and $g^t(\omega)$, respectively. The power spectrum $|\tilde{g}(\xi,\omega)|^2$ of $\tilde{g}(\xi,\omega)$, is given by

$$|\tilde{g}(\xi,\omega)|^2 = |\tilde{F}^a(\xi,\omega)|^2 + |\tilde{F}^b(\xi,\omega)|^2 = |\tilde{g}^x(\xi)|^2 * |\tilde{g}^t(\omega)|^2 \tag{31}$$

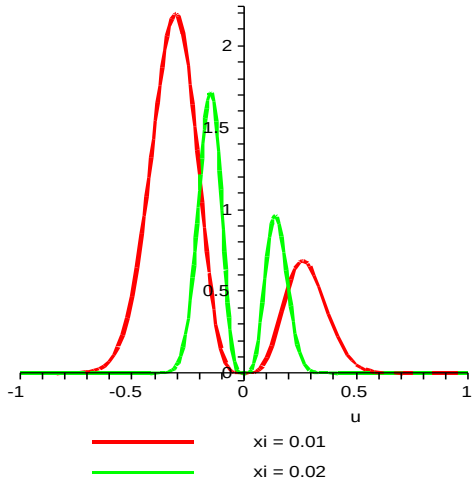

Figure 18: Curves showing the behaviour of $\tilde{g}^x(\xi)$ for two different frequencies f (0.01 and 0.02). In the graphs it is possible to see the maximal value reached along the negative spatial frequencies.

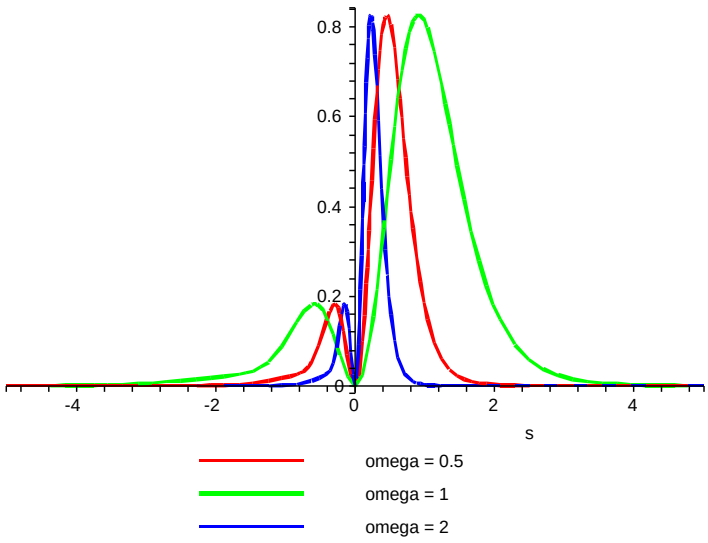

Figure 19: Curves showing the behaviour of $\tilde{g}^t(\omega)$ for three different values of τ (0.5, 1, 2). In the graphs it is possible to see the maximal value reached along the positive temporal frequencies.

Some graphs obtained for $\tilde{g}^x(\xi)$ and $\tilde{g}^t(\omega)$ can be seen in Figures 18 and 19, respectively.

Combining $\tilde{g}^t(\omega)$ and $\tilde{g}^x(\xi)$ it is possible to find frequency activity maps as the the ones shown in Figure 3 (*Right*).

The goal of this analysis is the bank filter design. To do so, it is required to find expressions relating the values of f, τ of equations (2) and (5) with the filter frequence response given by (31). These relationships were done using numerical analysis approximating the curves by polynomials of second order.

References

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2:284–299, 1985.

[2] EH Adelson and JA Movshon. Phenomenal coherence of moving visual patterns. *Nature*, 300(5892):523–525, 1982.

[3] J. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[4] J.A. Beintema and M. Lappe. Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences of the USA*, 99(8):5661–5663, 2002.

[5] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Proceeding IEEE International Conference on Computer Vision*, 2005.

[6] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. 23(3):257–267, March 2001.

[7] R. T. Born. Center-surround interactions in the middle temporal visual area of the owl monkey. *Journal of Neurophysioly*, 84:2658–2669, 2000.

[8] R.T. Born and D.C. Bradley. Structure and function of visual area MT. *Annu. Rev. Neurosci*, 28:157–189, 2005.

[9] David Bradley and Richard Andersen. Center-surround antagonism based on disparity in primate area mt. *Journal of Neuroscience*, 18(18):7552–7565, sep 1998.

[10] G. T. Buracas and T. D. Albright. Contribution of area mt to perception of three-dimensional shape: a computational study. *Vision Res*, 36(6):869–87, 1996.

[11] M Carandini, DJ Heeger, and JA Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Jounal of Neuroscience*, 17(21):8621–8644, nov 1997.

[12] A. Casile and M. Giese. Roles of motion and form in biological motion recognition. *Artifical Networks and Neural Information Processing, Lecture Notes in Computer Science 2714*, pages 854–862, 2003.

[13] A. Casile and M. Giese. Critical features for the recognition of biological motion. *Journal of Vision*, 5:348–360, 2005.

[14] J. Chey, S. Grossberg, and E. Mingolla. Neural dynamics of motion processing and speed discrimination. *Vision Res.*, 38:2769–2786, 1997.

[15] R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *5th Intl. Conf. on Automatic Face and Gesture Recognition*, 2002.

[16] B. Conway and M. Livingstone. Space-time maps and two-bar interactions of different classes of direction-selective cells in macaque v-1. *Journal of Neurophysiology*, 89:2726–2742, 2003.

[17] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. 22(8), August 2000.

[18] R De Valois, N. Cottaris, et al. Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity. *Vision Research*, 40:3685–3702, 2000.

[19] A. Delorme, L. Perrinet, and S. Thorpe. Network of integrate-and-fire neurons using rank order coding b: spike timing dependant plasticity and emergence of orientation selectivity. *Neurocomputing*, 38:539–545, 2001.

[20] A. Destexhe, M. Rudolph, and D. Paré. The high-conductance state of neocortical neurons in vivo. *Nature Reviews Neuroscience*, 4:739–751, 2003.

[21] M.-J. Escobar, A. Wohrer, P. Kornprobst, and T. Vieville. Biological motion recognition using an mt-like model. In *Proceedings of 3rd Latin American Robotic Symposium*, 2006.

[22] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, 1:1–47, 1991.

[23] D.M. Gavrila. The visual analysis of human movement: A survey. 73(1):82–98, 1999.

[24] W. Gerstner and W. Kistler. *Spiking Neuron Models*. Cambridge University Press, 2002.

[25] M.A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and actions. *Nature Reviews Neuroscience*, 4:179–192, 2003.

[26] L. Goncalves, E. DiBernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3D. In *5*, pages 764–770, Boston, MA, June 1995.

[27] E. Grossman, M.Donnelly, R. Price, D. Pickens, V. Morgan, G. Neighbor, and R. Blake. Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*, 12(5):711–720, 2000.

[28] Norberto Grzywacz and A.L. Yuille. A model for the estimate of local image velocity by cells on the visual cortex. *Proc R Soc Lond B Biol Sci.*, 239(1295):129–161, mar 1990.

[29] J. Hateren and A. van der Schaff. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings. Biological Science, Royal Society of London*, 265:359–366, 1998.

[30] D. Hogg. Model-based vision: a paradigm to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[31] D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat visual cortex. *J Physiol*, 160:106–154, 1962.

[32] H.E. Jones, K.L. Grieve, W. Wang, and A.M. Sillito. Surround suppression in primate v1. *Journal of Neurophysiology*, 86:2011–2028, 2001.

[33] L. L. Lui, J. A. Bourne, and M. G. P. Rosa. Spatial summation, end inhibition and side inhibition in the middle temporal visual area (mt). *Journal of Neurophysiology*, 97(2):1135, 2007.

[34] N. Majaj, M. Carandini, and Movshon J.A. Motion integration by neurons in macaque mt is local, not global. *The Journal of Neuroscience*, 27(2):366–370, jan 2007.

[35] D. R. Mestre, G. S. Masson, and L. S. Stone. Spatial scale of motion segmentation from speed cues. *Vision Research*, 41(21):2697–2713, September 2001.

[36] L. Michels, M.Lappe, and L.M. Vaina. Visual areas involved in the perception of human movement from dynamic analysis. *Brain Imaging*, 16(10):1037–1041, July 2005.

[37] J. A. Movshon and W. T. Newsome. Visual response properties of striate cortical neurons projecting to area mt in macaque monkeys. *Journal of Neuroscience*, 16(23):7733–7741, 1996.

[38] JA Movshon, EH Adelson, MS Gizzi, and WT Newsome. The analysis of moving visual patterns. *Experimental Brain Research*, 11:117–151, 1986.

[39] S. A. Niyogi and E. H. Adelson. Analyzing gait with spatiotemporal surfaces. *Vismod*, (290), 1994.

[40] C. C. Pack, J. N. Hunter, and R. T. Born. Contrast dependence of suppressive influences in cortical area mt of alert macaque. *Journal of Neurophysiology*, 93(3):1809–1815, Mar 2005.

[41] János Perge, Bart Borghuis, Roger Bours, Martin Lankheet, and Richard van Wezel. Temporal dynamics of direction tuning in motion-sensitive macaque area mt. *Journal of Neurophysiology*, 93:2194–2116, 2005.

[42] J. A. Perrone and R.J. Krauzlis. Motion integration by mt pattern neurons: An explanation for pattern-to-component effects. In *Perception 36 ECVP Abstract Supplement*, 2007.

[43] R. Polana and R.C. Nelson. Detection and recognition of periodic, non-rigid motion. *ijcv*, 23(3):261–282, 1997.

[44] Nicholas Priebe, Carlos Cassanello, and Stephen Lisberger. The neural representation of speed in macaque area mt/v5. *Journal of Neuroscience*, 23(13):5650–5661, jul 2003.

[45] D. Putthividhya and T.W. Lee. Motion patterns: High-level representation of natural video sequences. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[46] J.G. Robson. Spatial and temporal contrast-sensitivity functions of the visual system. *J. Opt. Soc. Am.*, 69:1141–1142, 1966.

[47] K. Rohr. Toward model-based recognition of human movements in image sequences. *CVGIP, Image Understanding*, 1:94–115, 1994.

[48] R. Van Rullen and S. Thorpe. Rate coding versus temporal order coding: What the retina ganglion cells tell the visual cortex. *Neural Computing*, 13(6):1255–1283, 2001.

[49] Alan Saul, Peter Carras, and Allen Humphrey. Temporal properties of inputs to direction-selective neurons in monkey v1. *Journal of Neurophysiology*, 94:282–294, 2005.

[50] M.P. Sceniak, M.J. Hawken, and R. Shapley. Visual spatial characterization of macaque v1 neurons. *Journal of Neurophysiology*, 85:1873–1887, 2001.

[51] S.M. Seitz and C.R. Dyer. View-invariant analysis of cyclic motion. 25(3), 1997.

[52] M. Shah and R. Jain. *Motion-based recognition*. Computational Imaging and Vision Series. Kluwer Academic Publisher, 1997.

[53] Rodrigo Sigala, Thomas Serre, Tomaso Poggio, and Martin Giese. Learning features of intermediate complexity for the recognition of biological motion. *ICANN 2005, LNCS 3696*, pages 241–246, 2005.

[54] E. P. Simoncelli and D.J. Heeger. A model of neuronal responses in visual area mt. *Vision Research*, 38:743–761, 1998.

[55] Matthew Smith, Najib Majaj, and Anthony Movshon. Dynamics of motion signaling by neurons in macaque area mt. *Nature Neuroscience*, 8(2):220–228, feb 2005.

[56] R. J. Snowden, S. Treue, R. G. Erickson, and R. A. Andersen. The response of area mt and v1 neurons to transparent motion. *The Journal of Neuroscience*, 11(9):2768–2785, Sep 1991.

[57] S. Thorpe, A. Delorme, and R. VanRullen. Spike based strategies for rapid processing. *Neural Networks*, 14:715–726, 2001.

[58] Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000.

[59] Liang Wang and David Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *Proceedings CVPR*, 2007.

[60] A.B. Watson and A.J. Ahumada. A look at motion in the frequency domain. *NASA Tech. Memo.*, 1983.

[61] D. J. Wielaard, M. Shelley, D. McLaughlin, and R. Shapley. How simple cells are made in a nonlinear network model of the visual cortex. *The Journal of Neuroscience*, 21(14):5203–5211, July 2001.

[62] HR Wilson, VP Ferrera, and C Yo. A psychophysically motivated model for two-dimensional motion perception. *Visual Neuroscience*, 9(1):79–97, jul 1992.

[63] D. Xiao, S. Raiguel, V. Marcar, J. Koenderink, and G. A. Orban. Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. *Proceedings of the National Academy of Sciences*, 92(24):11303–11306, 1995.

[64] D.-K. Xiao, V.L. Marcar, S.E. Raiguel, and Orban G.A. Selectivity of macaque mt/v5 neurons for surface orientation in depth specified by motion. *European Journal of Neuroscience*, 9:956–964, 1997.

[65] D. K. Xiao, S. Raiguel, V. Marcar, and G. A. Orban. The spatial distribution of the antagonistic surround of mt/v5 neurons. *Cereb Cortex*, 7(7):662–77, 1997.

[66] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proceedings of CVPR'01*, volume 2, pages 123–128, 2001.

[67] Lihi Zelnik-Manor and Michal Irani. Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1530–1535, sep 2006.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399